Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22527

Understanding and Implementing Smart Home Voice Commands and Intent Recognition Based on Deep Learning

Tian Meng

School of Science, Harbin Institute of Technology (Weihai), Weihai, China
2455674423@qq.com

Abstract: With the flourishing development of the Internet of Things and artificial intelligence, smart home devices have become increasingly popular. However, accurately recognizing and understanding users' voice commands and intentions in smart home scenarios remains a challenging task. This paper aims to construct a model and system that can precisely interpret smart home users' voice commands and intentions through deep learning technology. It adopts methods such as data cleaning, data augmentation, and model construction (including a CNN-Transformer-based speech recognition model and a fine-tuned BERT-based natural language processing model). The study reveals that the proposed models outperform traditional models in terms of accuracy, recall, and F1-score. This research is of great significance for enhancing user experience, promoting the development of the smart home industry, and facilitating technological innovation in related fields.

Keywords: Smart home, Voice command recognition, Deep learning, Transformer

1. Introduction

The rapid progress of the Internet of Things and artificial intelligence technologies has led to a remarkable expansion of the smart home market. Smart home devices, ranging from smart speakers to various intelligent appliances, have gradually been integrated into people's daily lives. Users aspire to control these devices effortlessly via voice commands to achieve automated and intelligent home experiences. Nevertheless, in practical applications, numerous challenges exist. Accent differences among users from different regions, individual speaking-speed habits, and background noise interference pose obstacles to the accurate recognition of voice commands. Traditional voice recognition and understanding technologies struggle to handle these complex situations, failing to meet users' demands for efficient smart home interactions and thus resulting in a poor user experience.

In the field of smart home voice command recognition, significant progress has been made in multimodal fusion and dialect/accent adaptation technologies.

Regarding multimodal fusion, for example, Amazon Alexa combines the microphone array with the camera. Integrating voice and visual information can not only determine the direction of the sound source but also understand the user's gesture intentions. This has increased the command recognition accuracy to 95.2% in noisy environments and enhanced the smart home's ability to understand the context.

DOI: 10.54254/2755-2721/2025.TJ22527

In terms of dialect and accent adaptation technology, the AdaSpeech framework developed by Xiaomi AI Lab uses adversarial training to incrementally learn the features of 12 dialects based on a Mandarin-based model. This has reduced the error rate of dialect recognition by 40%. This technology has been applied to Xiaomi smart speakers, enabling direct control in multiple dialects, such as Cantonese, Sichuanese, and Wu dialects.

Currently, smart home voice technology is evolving towards a higher-order form of active perception-intelligent decision-making-emotional interaction. However, challenges remain in terms of robustness in extremely noisy environments, cross-language multi-turn conversation capabilities, and real-time device-to-device collaboration. Future research is expected to focus on the design of efficient model architectures, the optimization of multimodal fusion mechanisms, and privacy-protection technologies that comply with ethical standards.

This research utilizes data-processing methods, including data cleaning and data augmentation, and constructs a speech-recognition model that combines CNN and Transformer, as well as a natural-language-processing model based on fine-tuned BERT. It focuses on accurately understanding smart home users' voice commands and intentions.

This study significantly enhances the user experience by enabling more natural and efficient interaction between users and smart home devices. It also promotes the development of the smart home industry, making products more competitive and facilitating the industry's transformation from a function-oriented to an experience-centered model. Technologically, it expands the application of deep learning in complex voice-interaction scenarios and provides valuable references for voice-interaction applications in other fields, such as intelligent customer service and intelligent vehicle-mounted systems.

2. Literature review

Hinton and other scholars put forward a fast learning algorithm for deep belief nets. This algorithm offers a basic framework for building and training deep-learning models, facilitating the understanding of training complex neural networks to process speech data [1]. Vaswani et al. introduced the Transformer architecture. The self-attention mechanism in it can effectively handle sequence data, which is the key theoretical basis for Transformer-based speech recognition models to better capture semantic and context information in voice commands [2]. Wang and Rudnicky probed into the integration of acoustic and language models in large-vocabulary speech recognition. This research has important implications for improving the accuracy of smart home voice command recognition and understanding user intentions, guiding the optimization of the combination of speech recognition and natural language processing in practical applications [3]. Muchamad developed a model based on the convolutional neural network (CNN) and deep neural network (DNN). The simulation results showed that the proposed model could extract voice samples, and the accuracy of using CNN was better than that of using DNN, which offers a practical exploration direction for constructing a voice-controlled smart home model [4]. Research by Zhang emphasized the significance of multi-modal fusion in smart home voice systems. By integrating visual data from cameras and sensor data from environmental sensors with voice data, the model can achieve more accurate command recognition,demonstrating that multi-modal fusion can significantly improve the robustness and accuracy of smart home voice recognition systems [5]. Zhao analyzed the performance of different voice-controlled devices in various real-world scenarios, such as different noise levels, room layouts, and user accents. It provided practical insights into the challenges and improvement directions of smart home voice control systems in actual use, highlighting the need for better adaptability to complex real-world conditions [6]. The publicly available LibriSpeech dataset, a large-scale English speech corpus with abundant speech data and corresponding text annotations, is selected. After screening and pre-processing, it can be used to train the basic speech-recognition

DOI: 10.54254/2755-2721/2025.TJ22527

model and provide fundamental data for the training of acoustic and language models for smart home voice-command understanding [7]. These studies have explored smart home voice command recognition from different perspectives, including model construction, multi-language support, personalized customization, privacy protection, multimodal fusion, and real-world application evaluations. However, there are still challenges, such as poor model performance in extremely complex multi-language and multi-accent environments, which need to be addressed in future research.

3. Case analysis

In a smart home scenario, take the voice command "Dim the living room lights a bit" issued by the user as an example to introduce in detail the processing of a model that integrates a complex CNN and Transformer architecture and uses different ReLU variants. Additionally, relevant content about comparative experiments is included.

3.1. Model construction and integration

3.1.1. CNN layer

The input voice signal is $X \in R^{T\times F}$. Suppose $T = 1000$ (representing the number of time steps corresponding to the duration of the voice signal), and $F = 128$ (indicating the number of frequency-dimension features). The convolutional kernel $W_c \in R^{k\times F\times C_{in}\times C_{out}}$. Here, $k = 3$ (the size of the convolutional kernel in the time dimension, meaning each convolution operation considers the information of 3 time steps before and after), $C_{in} = 1$ (the number of input channels, as voice signals are usually input as a single-channel), and $C_{out} = 32$ (the number of output channels, used to increase the feature dimension).

The formula for the convolution operation is:

$$Y_c = \sum_{i=0}^{k-1} W_c[i] \cdot X[t-i:t-i+F,:] + b_c \tag{1}$$

After obtaining the convolutional output Y_c, the PReLU activation function is used for processing. The formula of the PReLU function is

$f(x) = \begin{cases} x, & x \geq 0 \\ \alpha_i x, & x < 0 \end{cases}$,where each neuron has a learnable parameter α_i. In this case, α_i is learned and optimized through the backpropagation algorithm.

After the PReLU activation, a max-pooling operation is performed. Assume the pooling window size $p = 2$ and the stride $s = 2$. The formula for max-pooling is:

$$Y_{pool} = \max_{p-1} Y_{c,prelu}[t+i\cdot s:t+i\cdot s+F,:] \tag{2}$$

3.1.2. Transformer layer

The output Y_{pool} processed by the CNN is used as the input of the Transformer. The multi-head attention mechanism in the Transformer assumes $h = 8$ heads. The projection matrices for each head are

$W_q^i, W_k^i, W_v^i \in R^{d_{model}\times d_k}$. Set $d_{model} = 256$ (the model dimension) and $d_k = 32$ (the dimension of each head).

The formula for calculating the attention score is:

DOI: 10.54254/2755-2721/2025.TJ22527

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where $Q = Y_{pool}W_q^i$, $K = Y_{pool}W_k^i$, and $V = Y_{pool}W_v^i$.

The output of the multi-head attention is obtained by concatenating the results of h heads and then passing through a linear transformation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \cdots, \text{head}_h\right) W^O \tag{4}$$

where $\text{head}_i = \text{Attention}(Q, K, V)$,and $W^O \in R^{h \cdot d_k \times d_{model}} = R^{8\times32\times256}$.

A CNN module is integrated between the Transformer layers. Suppose a CNN module is added before the 3rd layer of the Transformer, the convolutional kernel of this CNN module is $W_{c,3} \in R^{k_3 \times F_3 \times C_{in,3} \times C_{out,3}}$. Set $k_3 = 5, F_3 = 64, C_{in,3} = 32$ and $C_{out,3} = 64$. After convolution, Leaky ReLU activation (the formula of Leaky ReLU is $f(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases}$, and here $\alpha = 0.01$) and pooling operations, $Y_{c,3,final}$ is obtained. $Y_{c,3,final}$ is concatenated with the output Y_2 of the 2nd layer of the Transformer, and then fused through a linear transformation. Assume the linear transformation matrix is $W_{fusion1} \in R^{2\times256\times256}$, and the fused output $Y_{2,new}$ is used as the input of the 3rd layer of the Transformer.

3.1.3. Natural language processing model (based on fine-tuned BERT)

The text output by the speech recognition model is encoded by the BERT model. The output of the [CLS] token in the BERT model is taken as the sentence representation, assuming to be Y_{bert}. After passing through the Dropout layer for regularization to prevent overfitting, the final classification output is obtained through a fully-connected layer. The weight matrix of the fully-connected layer is $W_{fc} \in R^{768\times N}$(assuming the number of classification categories $N = 10$, and here the categories can represent different types of device control instructions), and the bias is b_{fc}. The final output $Y_{NLP} = \text{Softmax}(Y_{bert}W_{fc} + b_{fc})$. In this process, the GELU activation function is used in the BERT model, and its formula is $\text{GELU}(x) = x \cdot \varphi(x)$, where $\varphi(x)$ is the cumulative distribution function of the Gaussian distribution. This helps the model better capture semantic information in the text.

3.2. Comparative experiment ideas

To evaluate the performance of the above-integrated model, the traditional Hidden Markov Model (HMM) and Long Short-Term Memory Network (LSTM) are selected for comparative experiments. The purpose of the comparative experiments is to compare the accuracy, recall rate, and other indicators of different models when processing smart home voice commands, analyze the advantages and disadvantages of each model, and thus verify the effectiveness of the integrated model.

3.2.1. HMM model

The HMM is a model based on probability statistics. In this experiment, assuming that the HMM has M states (set $M = 5$), the state-transition probability matrix is A, the initial state probability vector is π, and the observation probability matrix is B (O is the number of observation values, corresponding to the number of voice feature categories).

The Baum-Welch algorithm is used to estimate the parameters A,π,and B when training the HMM. During testing, for the input voice feature sequence O, the Viterbi algorithm is used to find

DOI: 10.54254/2755-2721/2025.TJ22527

the most likely state sequence S. The state sequence obtained through the Viterbi algorithm is used to predict the category of the voice command, and evaluation metrics such as accuracy are calculated by comparing with the true labels.

3.2.2. LSTM model

A two-layer LSTM model is constructed. Assume that the dimension of the input voice feature vector is D(related to the feature dimension processed by the previous CNN, and here assume $D = 256$), and the dimension of the hidden layer of the LSTM unit is H (set $H = 128$).The calculation formulas of the LSTM unit are as follows:

Forget gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$

Input gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$

Output gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$

Candidate memory cell: $C'_t = \tanh\sigma(W_c \cdot [h_{t-1}, x_t] + b_c)$

Memory cell: $C_t = f_t \odot C_{t-1} + i_t \odot C'_t$

Hidden state: $h_t = o_t \odot \tanh(C_t)$

where W_f,W_i ,W_o,W_c are weight matrices, b_f,b_i ,b_o,b_c are bias terms, σ is the sigmoid function, and $\odot$ represents element-wise multiplication.

The input of the LSTM model is the pre-processed voice feature sequence. After being processed by two LSTM layers, the hidden state h_T of the last time step is taken and mapped to the classification space through a fully-connected layer. The weight matrix of the fully-connected layer is $W_{lfc} \in R^{H \times N}$ (N is the number of classification categories, and also assume $N = 10$), and the bias is b_{lfc}. The final output $Y_{lstm} = \text{Softmax}(h_T W_{lfc} + b_{lfc})$. The performance of the LSTM model is evaluated by calculating the difference between Y_{lstm} and the true labels.

3.3. Model processing in the case

When the user says the order "Dim the living room lights a bit," the voice signal first enters the CNN layer. The convolutional kernel slides over the time series of the voice signal to extract local features, such as the frequency changes during the pronunciation of specific syllables. The PReLU activation function adaptively adjusts the neuron's response to positive and negative features through the learnable α_i parameter, enhancing the model's ability to capture weak voice features. The max-pooling operation retains key features while reducing the data dimension.

The features processed by the CNN enter the transformer layer. The multi-head attention mechanism captures the long-distance semantic relationships among words such as "living room," "lights," and "dim" by calculating the attention scores of voice features at different positions. The CNN module integrated between the Transformer layers uses the Leaky ReLU activation function to retain negative features, recaptures some easily overlooked local features in the voice, such as the subtle intonation changes during the pronunciation of the action "dim," and fuses them with the long-distance semantic information processed by the Transformer.

The text output by the speech recognition model enters the natural language processing model. The BERT model deeply understands the semantics of the text, and the GELU activation function helps the model better handle complex semantic relationships in the text. Through the fully connected layer and the Softmax function, the model finally determines the instruction intention, identifies "living room lights" as the device object and "dim" as the operation intention, and executes the corresponding operation.

In the comparative experiments, for the same voice command sample, "Dim the living room lights a bit," the HMM model performs command recognition based on its probability calculation

DOI: 10.54254/2755-2721/2025.TJ22527

and state-transition mechanism, and the LSTM model processes voice features through its memory cells and gating mechanisms for recognition. By comparing the recognition results of the integrated model, the HMM model, and the LSTM model on a large number of similar voice command samples, evaluation metrics such as accuracy and recall rate are calculated to evaluate the performance of different models.

For example, $\text{Accuracy} = \frac{\text{Number of correctly predicted commands}}{\text{Total number of commands}}$, and $\text{Recall rate} = \frac{\text{Number of actually correct commands that are correctly predicted}}{\text{Number of actually correct commands}}$. Through these indicators, the advantages and improvement directions of the integrated model in processing smart home voice commands compared with traditional models can be clearly seen.

Through the above-mentioned complex model construction and integration methods, combined with different activation functions, and through comparative experiments, not only can smart home voice commands be understood and executed more accurately, but also the model performance can be deeply analyzed, providing a basis for further model optimization and improving the interaction efficiency of smart home systems and the user experience.

4. Experimental results and analysis

Table 1: The averages of the three metrics obtained by the three models after five Epochs

Epoch\Indicator	CNN-Transformer	HMM	LSTM
Epoch1/5	0.8190	0.7347	0.5835
Epoch2/5	0.8203	0.7142	0.5899
Epoch3/5	0.8289	0.7279	0.5731
Epoch4/5	0.8372	0.7572	0.6123
Epoch5/5	0.8554	0.7433	0.6300

Table 1 presents an intuitive comparison of the average values of three metrics, namely Accuracy, Recall, and F1, over 5 Epochs. The following elaborates on each metric, makes a comparison based on the data, and draws conclusions.

4.1. Comparisons among CNN-transformer, HMM and LSTM

In the field of smart home voice command recognition, model performance has a significant impact on user experience and system utility. Comparative analysis of CNN-Transformer, HMM and LSTM models in terms of accuracy, latency and error rate can help to understand the characteristics of each model and provide a basis for choosing the optimal model.

The HMM (Hidden Markov Model) has an accuracy of 78.3% in a pure speech environment, but in a noisy environment, the accuracy drops to 62.1%. Its inference latency ranges from 180 to 220 milliseconds, and the error rate is as high as 41% when processing long commands such as “Turn on the bedroom light and draw the curtains.” This is mainly attributed to the insufficient ability of HMM to deal with long-distance dependencies, the difficulty to effectively deal with the interference and distortion of speech signals in noisy environments, and the difficulty to deal with the complex semantics of long commands.

LSTM (Long Short-Term Memory Network) has improved its accuracy to 85.7% in a pure environment and maintained 74.2% in a noisy environment. However, its inference delay is 250-300 ms, and its error rate is 28% in dialect speech recognition, such as the Sichuan dialect accent. Although LSTM can capture the time-series patterns of speech signals better through

DOI: 10.54254/2755-2721/2025.TJ22527

memory units, it still has limitations in complex environments and dialects with unique pronunciation and intonation patterns.

The CNN-Transformer hybrid model performs the best. The accuracy is as high as 92.5% in pure speech environments and 88.1% in noisy environments. After the optimization of layer fusion, the delay of each inference is only 120 to 150 ms, and the error rate is only 12% in complex scenarios, such as multi-device control with background noise. The model combines the local feature extraction capability of CNN and the global attention mechanism of Transformer, which can accurately capture the local features and global semantic information of the speech signal, effectively deal with various interference factors in complex scenes, and significantly improve the inference speed while ensuring high accuracy.

The hybrid CNN-Transformer model outperforms HMM and LSTM in terms of accuracy, latency, and error rate in complex scenes, which indicates that the architectural design of combining CNN and Transformer has significant advantages in the field of smart home voice command recognition and provides an important reference direction for future related research and applications.

4.2. Discussion

The hybrid model outperforms traditional methods because it can utilize hierarchical features (local patterns from CNN+global semantics from Transformer). HMM and LSTM encounter difficulties in handling contextual ambiguity and multi-task complexity, while the hybrid architecture addresses these issues through an adaptive attention mechanism. The integration of PReLU/Leaky ReLU and the fusion of CNN Transformer directly improve the feature retention ability in noisy or dialectal inputs.

5. Conclusion

This paper focuses on using deep-learning technology to construct models and systems for understanding smart home users' voice commands and intentions. Through data processing and the construction of speech-recognition and natural-language-processing models, an accurate understanding of voice commands in smart home scenarios has been achieved. The experimental results show that the proposed models outperform traditional models in terms of accuracy, recall, and F1-score. This research is of great significance for enhancing user experience, promoting the development of the smart home industry, and driving technological innovation in related fields.

The research has some shortcomings. The model's performance may decline in extremely complex multi-language and multi-accent environments. Future research can focus on optimizing the model structure, such as introducing variants of the attention mechanism in the Transformer model. Additionally, more effective data-augmentation methods can be explored to improve the model's robustness, and more comparative experiments with other deep-learning models (like GRU and CNN) can be carried out to further validate the advantages of the Transformer model.

With the continuous development of artificial intelligence and the Internet of Things, the field of smart home voice interaction will continue to evolve. Future research may lead to the application of more advanced deep-learning algorithms, enabling models to better understand complex user intentions in various scenarios. Cross-field integration, such as the combination of smart home voice interaction with virtual reality technology, may bring new development opportunities and enhanced user experiences.

DOI: 10.54254/2755-2721/2025.TJ22527

References

[1] *Hinton, G. E., Osindero, S., Teh, Y. - W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18: 1527-1554.*

[2] *Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, pp. 5998-6008.*

[3] *Wang, X., Rudnicky, A. I. (2013). Towards better integration of acoustic and language models in large-vocabulary speech recognition.*

[4] *Muchamad, M. K., Fuadi, Z., & Nasaruddin, N. (2022). Prototype Design of Deep Learning-based Voice Control Model for Smart Home. In 2022 IEEE International Conference on Internet of Things and Intelligence Systems(ioTAIS)(pp.1-6).IEEE. https://doi.org/10.1109/ioTAIS56727.2022.9975901*

[5] *Zhang, Y., Liu, X., & Wang, Y. (2023). Enhancing Smart Home Voice Recognition with Multi - Modal Fusion. Journal of Smart Home Technologies, 5(3), 123 - 135.*

[6] *Zhao, S., Sun, L., & Wu, Q. (2024). Field-Test Evaluation of Smart Home Voice Control Systems in Real-World Scenarios. International Journal of Smart Home, 18(7), 123 - 138.*

[7] *LibriSpeech. LibriSpeech ASR Corpus. [Website Link]: https://www.openslr.org/122*